# MuS: Multilingual Synergy with Shared Representations for Visual Speech Recognition

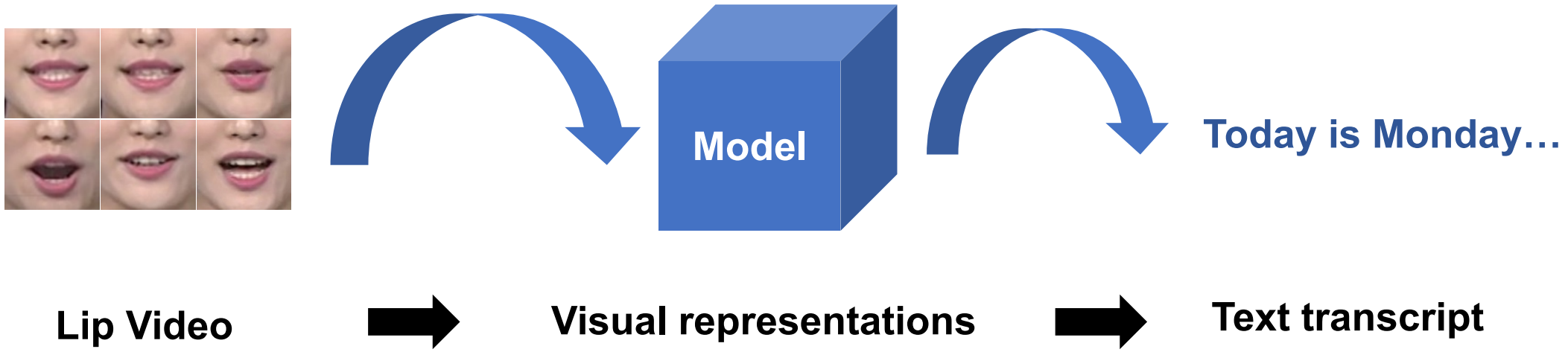
Yuheng Fan[1,2] Shuang Yang[1,2] Shiguang Shan[1,2] Xilin Chen[1,2]

[1]Institute of Computing Technology, Chinese Academy of Sciences

[2]University of Chinese Academy of Sciences

# Introduction

## Visual Speech Recognition(VSR): Reading Speech from Lip Movements




- Speech recognition in noisy environments
- Privacy-sensitive and silent-interaction scenarios

# Introduction

## From Monolingual to Multilingual VSR

Why Multilingual?

### Limited per-language

Video data are limited for any single language

### Scalability

One unified model for multiple languages

### Shared articulation

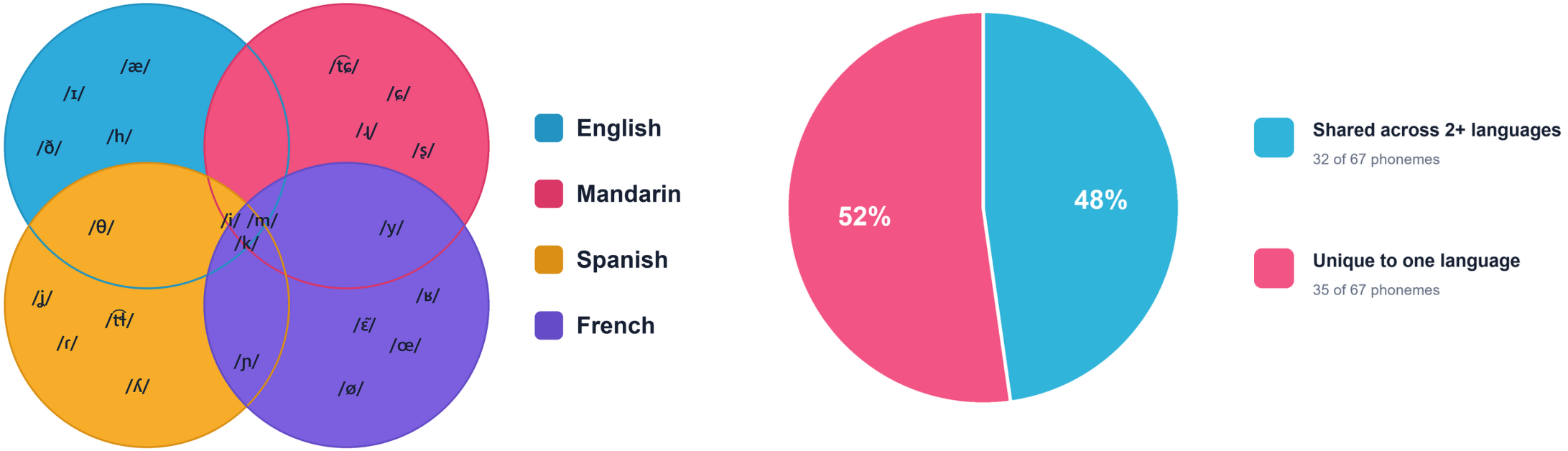
Different languages may share phonemes and corresponding lip-movement patterns

**Joint training ≠ joint gains**


**The “curse of multilinguality”**

# Motivation

## Shared Phonetic Patterns




Cross-Lingual Phoneme Overlap

# Motivation

## Shared Linguistic Structures

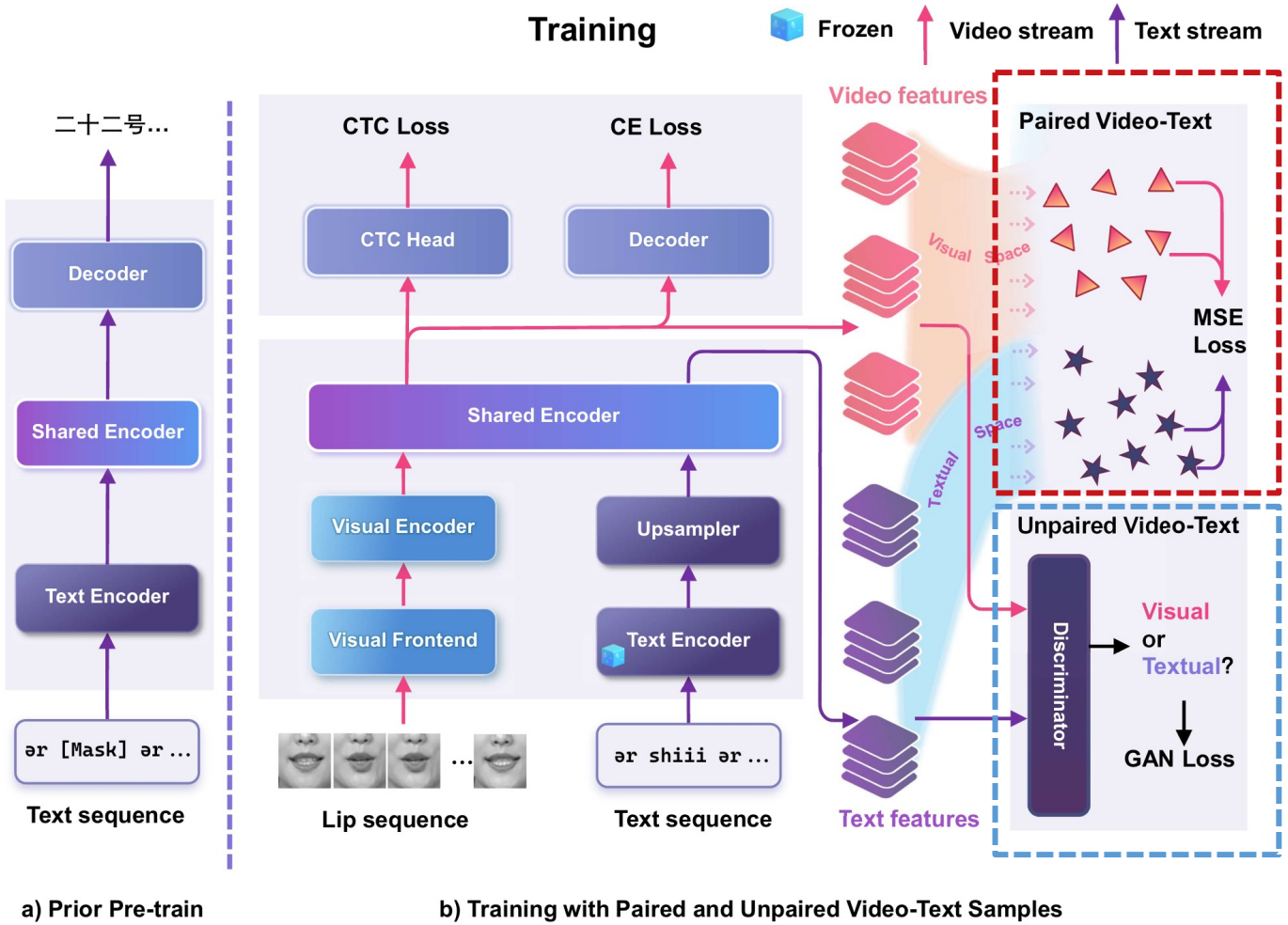
Different languages exhibit similar word-order and sentence patterns.

eg.


Contextual and syntactic regularities can provide transferable linguistic cues.

# Methodology

## Overall Framework

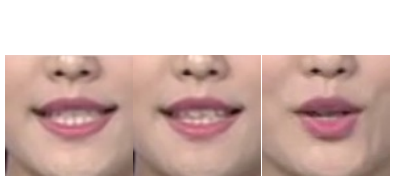



a) Linguistic prior pre-training

b) Paired & Unpaired Learning

# Methodology

## Paired Samples: MSE Loss



Visual Encoder

Upsampled Phoneme Sequence

Text Encoder

Shared Encoder

Video feature (Hv)

Phoneme feature(Hp)

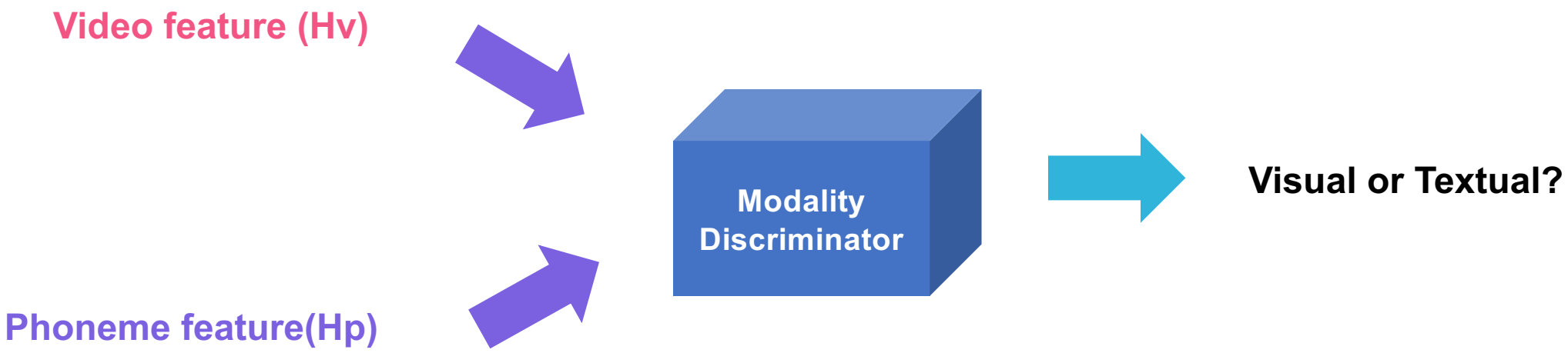
$L_{\text{MSE}}$

$$L_{\text{MSE}} = \frac{1}{|B|} \sum_{b=1}^{|B|} \left|\left| H_v - H_p \right|\right|_2^2$$

Same utterance → direct visual–phoneme representation matching

# Methodology

## Unpaired Samples: Discriminate Loss




$$L_{\mathrm{u}} = \min_{M} \max_{C} E_{H_p}[\log D(H_P)] + E_{H_v}[\log(1 - D(H_v))]$$

No content correspondence → align distributions, not individual instances

# Experiments

## Consistent Performance Improvements Across All Languages

**Bilingual**

**Table 1.** Bilingual Experimental Results

| Method | Modality | WER (%) | |
|---|---|---|---|
| | | en | zh |
| *Monolingual* | | | |
| BaseModel (en) | V | 37.55 | – |
| BaseModel (zh) | V | – | 60.95 |
| *Bilingual* | | | |
| BaseModel | V | 41.77 | 58.86 |
| **MuS (Ours)** | **V + T** | **31.47** | **47.91** |

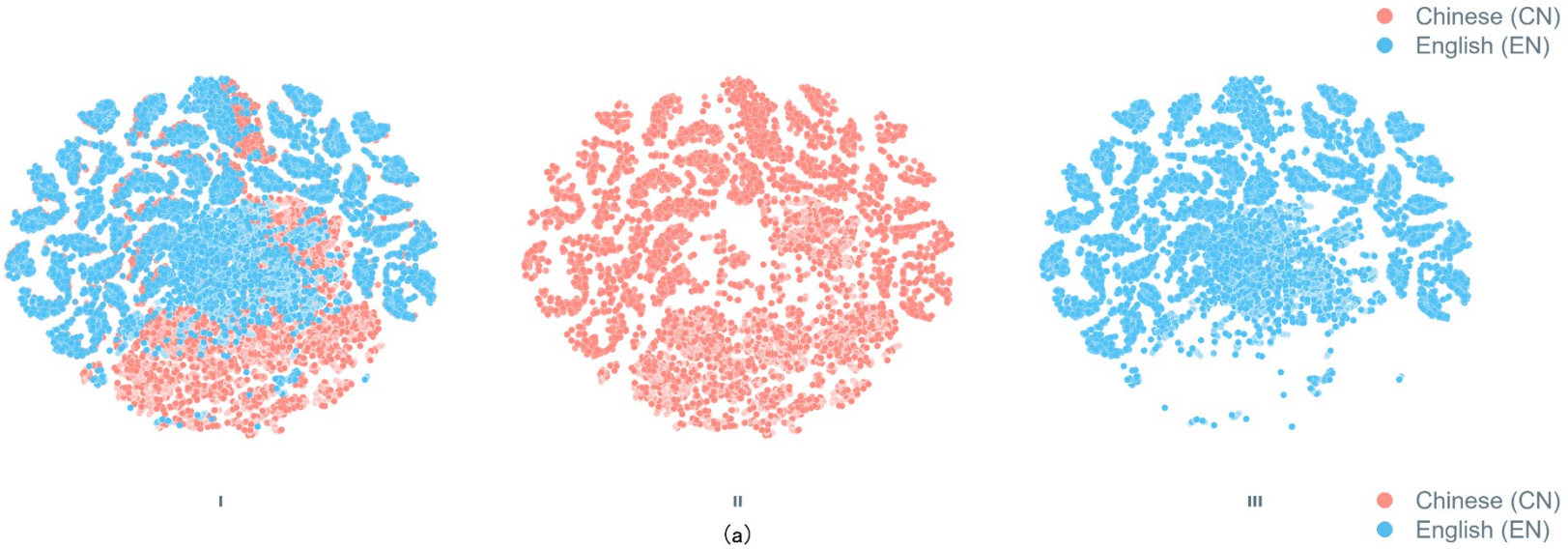
**Multilingual**

**Table 4.** Evaluation of Multilingual Visual Speech Recognition Methods.

| Method | Pre-train Data | Training Data | Single Model | WER (%) | | | |
|---|---|---|---|---|---|---|---|
| | | | | en | zh | es | fr |
| BaseModel | – | 542h | ✓ | 40.04 | 56.93 | 76.93 | 83.26 |
| **MuS (Ours)** | – | 542h | ✓ | **31.44** | **47.20** | 65.89 | 75.88 |

# Experiments

## Visualizing Learned Video Representations




**Base model**
**Looser, more language-specific clusters**

**MuS (Ours)**
**More compact and language-invariant**

# Thanks for listening!

fanyuheng25@mails.ucas.ac.cn